10/039,849

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００１年　９月　４日

出願番号
Application Number: 特願２００１－２６６６２９

[ST. 10/C]: [ＪＰ２００１－２６６６２９]

願人
Applicant(s): 株式会社日立製作所

２００４年１２月２２日

特許庁長官
Commissioner,
Japan Patent Office

小川

CERTIFIED COPY OF
PRIORITY DOCUMENT

出証番号　出証特２００４－３１１７３１５

【書類名】　特許願
【整理番号】　K01000441A
【あて先】　特許庁長官殿
【国際特許分類】　G06F　3/06
【発明者】
　【住所又は居所】　神奈川県小田原市中里３２２番地２号　株式会社日立製作所　ＳＡＮソリューション事業部内
　【氏名】　日野　直樹
【発明者】
　【住所又は居所】　神奈川県小田原市中里３２２番地２号　株式会社日立製作所　ＲＡＩＤシステム事業部内
　【氏名】　占部　喜一郎
【発明者】
　【住所又は居所】　神奈川県小田原市中里３２２番地２号　株式会社日立製作所　ＲＡＩＤシステム事業部内
　【氏名】　中野　俊夫
【特許出願人】
　【識別番号】　000005108
　【氏名又は名称】　株式会社日立製作所
【代理人】
　【識別番号】　100075096
　【弁理士】
　【氏名又は名称】　作田　康夫
【手数料の表示】
　【予納台帳番号】　013088
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】明細書

【発明の名称】データ格納システム

【特許請求の範囲】

【特許請求の範囲】

【請求項１】

分散して設置されたデータを格納する第一の記憶装置と第二の記憶装置間でリモートコピー機能により作成されたペアボリュームを、２台以上の上位装置から共用するデータ格納システムにおいて、１台の前記上位装置が前記ペアボリュームを排他的に占有し、他の前記上位装置からの更新要求を拒否する制御用ペアボリュームと、データ格納用ペアボリュームから成り、前記制御用ペアボリュームに２台以上の前記上位装置から更新要求があった場合に、前記制御用ペアボリュームに最も早くアクセスを行った前記上位装置が前記制御用ペアボリュームを排他的に使用する権利を獲得するデータ格納システム。

【請求項２】

請求項１記載のデータ格納システムにおいて、ペアボリュームの属性及び、前記ペアボリュームの属性を遷移させるペアボリューム制御コマンドを発行する、ペアボリューム制御用ソフトウェアを使用することで、前記制御用ペアボリュームに対して更新要求を発行できる上位装置を排他的に決定するデータ格納システム。

【請求項３】

請求項１または請求項２のデータ格納システムにおいて、上位装置上の汎用アプリケーションと連携して動作し、前記記憶装置に対し前記ペアボリューム制御コマンドを発行する前記ペアボリューム制御用ソフトウェアを使用する事で、前記制御用ペアボリュームに対して更新要求を発行できる上位装置を排他的に決定するデータ格納システム。

【請求項４】

２台以上の上位装置と、２台の記憶装置がデータ転送インターフェース及びネットワークインターフェースで接続されるクラスタリングシステムにおいて、前記上位装置もしくはネットワークインターフェースの障害発生時に前記上位装置

の専用使用権を一意に決定するための前記記憶装置内の判定用ボリュームをリモートコピー機能で二重化する、データ格納システム。

【請求項５】

２台以上の上位装置と、２台の記憶装置がデータ転送インターフェース及びネットワークインターフェースで接続されるクラスタリングシステムにおいて、前記上位装置もしくはネットワークインターフェースの障害発生時に前記上位装置の専用使用権を一意に決定するための前記記憶装置内の判定用ボリュームをリモートコピー機能で二重化し、複数の前記上位装置から前記判定用ペアボリュームを１つのボリュームとして認識するクラスタリングシステムを実現するデータ格納システム。

【請求項６】

分散して設置されたデータを格納する第一の記憶装置と第二の記憶装置間でリモートコピー機能により作成されたペアボリュームを、２台以上の上位装置から共用するデータ格納システムにおいて、ペアボリュームの専有使用を目的として前記上位装置から発行されたＳＣＳＩリザーブコマンドを前記ペアボリューム制御用ソフトウェアを介して、前記記憶装置内のボリュームに発行し、前記ＳＣＳＩリザーブコマンドに対する前記ボリュームから正常終了のレスポンスを前記ペアボリューム制御用ソフトウェアがうけて、ペアボリューム制御コマンドを前記ボリュームに発行し、前記ボリュームの属性の確認及び、前記属性に応じて、前記ボリュームの前記属性を遷移させる前記コマンドを発行し、正常終了した場合のみ、前記ＳＣＳＩリザーブコマンドに対する正常終了を前記上位装置に応答し、前記ボリュームに対して更新要求を発行できる上位装置を排他的に決定するデータ格納システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は分散して設置された２台以上の上位装置と２台の記憶装置から構成されるデータ格納技術に関する。

【０００２】

更に具体的には分散して設置された２台のリモートコピー機能を持つ記憶装置間で作成された2重化ボリュームの排他的制御を行うデータ格納技術に関する。

【０００３】

【従来の技術】

システムを止めることなく24時間サービスを継続させるためのさまざまなIT技術がいま、特にEビジネス化を推進する企業の関心を集めている。その中で高可用性を実現する技術の一つがクラスタリングである。複数台のシステムを組み合わせて，一部システムの障害時にもシステム全体が停止せず稼動できるように構築することで、データベース業務を行うシステム等で主に使われている。例えば、小規模システムの例では数台のサーバーで１つの共有ディスクを利用する構成をとり、サーバーと共有ディスク間はＳＣＳＩインターフェースを用い、この構成で数台のサーバーが扱っているデータを共有ディスクに置いておけば、一台のサーバーがダウンしても、他の一台のサーバーが共有ディスクのデータを引き継ぎ、処理を継続することができる。これをフェイルオーバという。このようにフェイルオーバを行うクラスタリングシステムでは、クライアントからはただ一つのサーバーを引き続き利用しているように業務を行える。

【０００４】

また、MicroSoft社のクラスタリング製品であるＭＳＣＳ（マイクロソフト・クラスタ・サーバ）ではクラスタリングシステム内でフェイルオーバを行う上位装置を一意に決定する為に、クラスタ構成情報を管理するクォーラムリソースと呼ばれる一つの判定用ボリュームを持つ。この判定用ボリュームに対して排他制御を行うことで、そのデータの一貫性を維持できる。

【０００５】

すなわち、複数の異なるプロセスが同時に一つのリソースに対して重複処理をしてしまうことにより、データの一貫性が損なわれるという問題を回避している。

【０００６】

次に、２台以上の上位装置と、２台の記憶装置がＳＣＳＩインターフェースによって接続される場合のクラスタリングシステムを例として説明する。クラスタ

リングシステムでは、各々の上位装置間をネットワークインターフェースで接続し、各上位装置は互いにメッセージを発信してハートビート通信を行う事で、お互いの稼動状況を監視する。

【０００７】

さらに、クラスタリングシステムは上位装置の障害又はネットワークインターフェースの障害が発生し、ハートビート通信が不可能になったことを検出して、個々の上位装置が他の上位装置のリソース、アプリケーション及びサービスの引き継ぎを行うかどうかの判断を行う。その判断を行う唯一の決定要素としてクラスタ構成情報を管理する１つの判定用ボリュームを記憶装置内に設ける。

【０００８】

一般にＳＣＳＩインターフェースを使用するシステムにおいては、複数のホストのうちひとつのホストがひとつのターゲット（例えば、磁気ディスクドライブ）を排他制御（専用使用）をする場合、最初にＳＣＳＩリザーブコマンドを発行することにより、そのターゲットをリザーブでき専用使用が可能となる。

【０００９】

上位装置の障害又はネットワークの障害等が発生し、ハートビート通信が不可能になった場合、各々の上位装置はクラスタリングシステム内での専用使用権を獲得する為に、判定用ボリュームに対してＳＣＳＩリザーブコマンドを発行し、専用使用権を手に入れることを試みる。判定用ボリュームへのＳＣＳＩリザーブコマンドの実行が成功した上位装置は他の上位装置のアプリケーション、サービス及びディスクボリューム等のリソースを全て引き継ぎ、フェイルオーバを実行する。一方、判定用ボリュームに対するＳＣＳＩリザーブコマンド実行が失敗した上位装置はその上位装置上で実行していたアプリケーション及びサービスを全て停止する。

【００１０】

このように、クラスタリングシステムは、一台の上位装置がクラスタ構成情報を管理する判定用ボリュームを専用使用することで実現されている。

【００１１】

【発明が解決しようとする課題】

特に大量のデータをリアルタイムで処理する必要がある、大規模なオンライントランザクションシステムにおいてはシステムを止めないために、サーバーをはじめディスク装置や電源などを多重化し、一つが故障しても他方が処理を引き継ぐ、フェイルオーバ機能をサポートしたクラスタリングシステムが必要になる。

【００１２】

従来技術によれば、クラスタリングシステムにおける判定用ボリュームは全ての上位装置がアクセスすることができる唯一、１つのディスクボリュームによって構成されている。

【００１３】

しかし、複数の上位装置が分散配置されるクラスタリングシステムにおいては、１つの判定用ボリュームを全ての上位装置がアクセスできるように設置するには、上位装置と判定用ボリュームを有する記憶装置間のインターフェース、例えばＳＣＳＩインターフェース等、接続用のケーブルの接続距離の制限がありシステム構築が出来ない場合があるという問題があった。

【００１４】

これを解決する手段として、災害バックアップシステムや遠隔地へのデータ送信システムを構築できるリモートコピー機能をクラスタリングシステムに用いることが考えられる。

【００１５】

リモートコピー機能では上位装置を経由せずに（ホストフリー），記憶装置間のリモートコピーを実現する。リモート側への更新データ反映順序を守るため，リモート側のコピーデータの論理的一貫性の保証が可能である。ここで、リモートコピー機能により分散配置された記憶装置間で二重化されたペアボリューム（それぞれを正ボリューム及び副ボリュームと呼び、正ボリュームを有する記憶装置及び副ボリュームを有する記憶装置がある）を作成し、ペアボリュームの正ボリュームもしくは副ボリュームは同一のデータを持つことが保証されているので、ある一台の上位装置のみが正ボリュームと副ボリュームのどちらかを判定用ボリュームとして専用使用する上位装置を決定する方法が考えられる。

【００１６】

しかし、分散配置された各システムはそれぞれ、そのシステム内でそれぞれのＳＣＳＩインターフェースを使用しているシステムゆえ、各上位装置はそれぞれのシステム内でＳＣＳＩリザーブコマンドを発行し、実行が成功した上位装置がボリュームを排他制御できることになり、正ボリュームを専用使用する上位装置及び副ボリュームを専用使用する2台の上位装置が存在することになる。

【００１７】

すなわち、正ボリュームもしくは副ボリュームを排他的に使用する目的でＳＣＳＩリザーブコマンドを発行しただけでは、正ボリュームもしくは副ボリュームは別々にリザーブすることができる為に、判定用ボリュームとして排他使用ができないという問題があった。

【００１８】

【課題を解決するための手段】

前述に示す通り、リモートコピー機能を用いて判定用ボリュームを設定する場合、ペアを構成する正ボリューム及び副ボリュームは各上位装置から別々のボリュームとして認識され、各上位装置が正ボリューム及び副ボリュームに対し、ＳＣＳＩリザーブコマンドを発行した場合、正ボリューム及び副ボリュームを個々にリザーブすることができ、結果として複数の上位装置が同じアプリケーション及びサービスを同時に稼動してしまい、クラスタリングシステムとしての機能をはたさないことになる。

【００１９】

従って、リモートコピー機能をもつ２台の記憶装置間で作成されたペアボリュームを、複数の上位装置からなるクラスタリングシステムにおいて１つの判定用ボリュームとして認識させ、クラスタリングシステム内の複数の上位装置間で当該ペアボリュームに対し更新要求を発行できる上位装置を排他的に決定する手法を実現する為に、本発明では上位装置上のクラスタソフトウェアから発行されるＳＣＳＩコマンドのリザーブコマンドを、ペアボリュームの「状態」（属性）を操作するペアボリューム制御用ソフトウェアが一旦受け、ペアボリューム制御用ソフトウェアが発行するペアボリューム制御コマンドと組み合わせて記憶装置に発行し、そのコマンドに対する記憶装置のレスポンスが、最初にリザーブコマン

ドを発行した上位装置のみ正常終了とすることで、ペアボリュームを排他的使用する上位装置を一意に決定することが可能となるデータ格納システムを提案する。

【００２０】

【発明の実施の形態】

以下、本発明の一実施例を図面を用いて説明する。尚、クラスタリングシステムを利用した実施例としたが、本発明の適用範囲はクラスタリングシステムに限ったものではない。

また、２台の上位装置からなるクラスタリングシステムを実施例として用いたが、上位装置は３台以上あっても構わない。

【００２１】

図１は本発明の一実施形態である。図１はそれぞれ上位装置と記憶装置からなる２つのサイト１００、１０１をネットワークインターフェースによって接続し、クラスタリングシステムを実現する構成図である。１０、１１はＣＰＵを持つ上位装置であり、上位装置１０、１１上ではクラスタソフトウェア１２，１３及び本発明で提案するペアボリューム制御用ソフトウェア８４，８５等が動作する。また、上位装置１０，１１はハートビート用ネットワーク５０によって接続され、定期的にお互いの稼動状況を監視する。

【００２２】

２０、２１は上位装置１０、１１からのデータを格納しておく記憶装置であり、データ転送インターフェース３０、３１を介して接続される。また、記憶装置２０、２１は複数のボリュームから構成され、上位装置１０、１１上にインストールされたアプリケーションからのデータ更新及び参照要求を受けることができる。

【００２３】

さらに、記憶装置２０、２１はリモートコピー機能を持ち、一方の記憶装置上の正ボリューム（Ｐ－ＶＯＬ）２２のミラーをもう一方の記憶装置上の副ボリューム（Ｓ－ＶＯＬ）２３に構築することができる。上位装置１０で動作するアプリケーションは記憶装置２０内の正ボリューム（Ｐ－ＶＯＬ）２２に対してデー

タの更新要求を行い、その更新内容がリモートコピー用インターフェース４０を通じて自動的にもう一方の記憶装置内の副ボリューム（Ｓ－ＶＯＬ）２３に反映される。

【００２４】

図２はクラスタリングシステムの内部構造を示す。上位装置１０，１１はそれぞれクラスタソフトウェア８０、８１、アプリケーション８２，８３及びペアボリューム制御用ソフトウェア８４，８５から構成される。ペアボリューム制御用ソフトウェア８４、８５はクラスタソフトウェア８０，８１と標準デバイスドライバ９０，９１の中間に位置する。

【００２５】

クラスタソフトウェア８０，８１はペアボリューム制御用ソフトウェア８４、８５が独自に定義したデバイスファイルに対してＳＣＳＩコマンドを発行し、それを受け取ったペアボリューム制御用ソフトウェア８４、８５は、必要に応じてペアボリューム制御コマンドと組み合わせたコマンドを標準デバイスドライバ９０，９１経由でクラスタ構成情報を管理する判定用ペアボリューム６０，６１へ発行する。

【００２６】

また、記憶装置２０，２１はクラスタの専有使用権を確保する為に構築された判定用ペアボリューム６０，６１、及びアプリケーション用のデータを格納するデータボリューム７０－７３で構成される。リモートコピー機能によって各ボリュームは二重化されており、リモートコピー機能の仕様により、各アプリケーションは正ボリューム（Ｐ－ＶＯＬ）のみに更新要求を発行する事ができる。

【００２７】

クラスタソフトウェア８０と８１はネットワークを介してハートビート通信を行うことで定期的に相手側上位装置の稼動状況を監視する。上位装置１０又は１１の障害、又はハートビートネットワーク５０（図１）の障害によりハートビート通信が不可能となった場合、クラスタソフトウェア８０、８１はペアボリューム制御用ソフトウェア８４，８５を介して、記憶装置２０，２１内に構築された判定用ペアボリューム６０，６１にそれぞれＳＣＳＩリザーブコマンドを発行す

る。

【００２８】

そして、判定用ペアボリュームをリザーブすることに成功した上位装置はクラスタソフトウェアによりクラスタリングシステム内に登録された全リソース及びアプリケーションの引き継ぎを行う。判定用ペアボリュームをリザーブすることに失敗した他の上位装置はクラスタソフトウェアによりリソースを解放し、アプリケーションを停止する。

【００２９】

ここで、図３に示すペアボリュームの状態遷移について詳細説明する。リモートコピー機能によって生成された各ペアボリュームは一般的に「非ペア状態」、「コピー状態」「ペア状態」、「サスペンド状態（正常）」、「サスペンド状態（異常）」の状態を持ち、上位装置上にインストールされたペアボリューム制御用ソフトウェアがペアボリューム制御コマンドを記憶装置へ発行することで、リモートコピー用インターフェース４０を通じて自動的に各々のペアボリュームの状態を遷移させることができる。

【００３０】

「非ペア状態」は２つのボリューム間にペアが形成されてない状態、「コピー状態」は２つのボリューム間で正ボリュームから副ボリュームへコピー実行中でペアボリュームのデータはまだ全て同等ではない状態。コピー完了時、ボリュームは「ペア状態」に変わる

「ペア状態」では、上位装置から正ボリュームへの全ての更新は副ボリュームに対して二重化される。「サスペンド状態（正常）」は正ボリュームに対する更新を副ボリュームに反映しない状態。例えばペアボリューム間で正ボリュームのみ更新をかけたい時、正、副ボリュームを本状態に設定する。

【００３１】

「サスペンド状態（異常）」は前述と違い、何らかの障害検知で、ペアボリューム間で内容の更新を保持できない時、正、副ボリュームを本状態に設定する。各ペアボリュームの状態を遷移させるペアボリューム制御コマンドは正ボリュームまたは副ボリュームのどちらに接続された上位装置からでも実行可能である。

【００３２】

例えば、ペアボリュームが「サスペンド状態（正常）」であった場合、正ボリューム側の上位装置からペア再同期コマンドを発行し「ペア状態」に遷移させる事も可能であるし、副ボリューム側の上位装置からペア再同期コマンドを発行し、「ペア状態」に遷移させることもできる。さらに、副ボリューム側の上位装置からペアスワップ再同期コマンドを発行し、副ボリュームを正ボリュームにスワップした上で「ペア状態」に遷移させることもできる。

【００３３】

また、ペアボリューム制御コマンドの実行結果は実行先のペアボリュームの状態に依存するという性質を持つ。以下詳細説明する。

【００３４】

図４にペアボリューム制御コマンドであるペア再同期コマンド及びペアスワップ再同期コマンド発行時の記憶装置からのレスポンスの例を示す。例えば、実行コマンドとしてペア再同期コマンドを発行した場合、実行先のボリュームが正ボリュームでかつ「サスペンド状態（正常）」でなければ記憶装置からのレスポンスは正常終了とはならない。

【００３５】

もし、実行先のボリュームが「コピー状態」もしくは「ペア状態」であった場合はそのペア再同期コマンドは異常終了する。また、ペアスワップ再同期コマンドは実行先のペアボリュームが副ボリュームであり、かつ「サスペンド状態（正常）」でなければ記憶装置からのレスポンスは正常終了とはならない。

【００３６】

もし、実行先のボリュームが「コピー状態」もしくは「ペア状態」であった場合はそのコマンドは異常終了する。このペアボリュームの性質を用いて、ペアボリューム制御用ソフトウェアはクラスタソフトウェアから発行されたＳＣＳＩコマンドを一旦受け取り、本ＳＣＳＩコマンドを分析した結果、リザーブコマンド等のペアボリューム間で排他的な処理を必要とするＳＣＳＩコマンドであった場合はペアボリューム制御コマンド例えば、ペア再同期コマンドと組み合わせて記憶装置側へ発行する。

【００３７】

そして、ペア再同期コマンドが記憶装置内で実行され、正常にペアボリューム状態を遷移できた場合は、正常終了ステータスをペアボリューム制御用ソフトウェアに送信する。さらに、正常終了ステータスを受け取ったペアボリューム制御ソフトウェアはクラスタソフトウェアに対して正常終了ステータスを返す。

【００３８】

正常終了ステータスを受け取ったクラスタソフトウェアはクラスタリングシステム内で引き続き生存することが許可され、フェイルオーバを実行する。一方、１つの上位装置から最初に発行されたペア再同期コマンドが記憶装置上で実行され、ペアボリュームの状態が「ペア状態」に遷移したことによって、他の上位装置から同様のペアボリューム制御コマンドが発行されたとしても図４に示すよう、「ペア状態」に発行されたペアボリューム制御コマンドは全て異常終了となる。

【００３９】

ペアボリューム制御コマンドの異常終了報告を受けたペアボリューム制御用ソフトウェアはＳＣＳＩコマンド要求が失敗したことの報告として、クラスタソフトウェアへ異常終了ステータスを返す。異常終了ステータスを受け取ったクラスタソフトウェアはクラスタリングシステム内での生存が許されず、担当しているアプリケーション及びサービスを全て停止させる。

【００４０】

図５、６で、本発明で提案するペアボリューム制御用ソフトウェアの動作についてＳＣＳＩリザーブコマンドを受けた場合とそれ以外のコマンドをうけた場合について更に詳細に説明する。ペアボリューム制御用ソフトウェア８４はクラスタソフトウェア８０から発行されたＳＣＳＩコマンドがリザーブコマンドであった場合（１１０）、判定用ペアボリュームの正ボリューム（Ｐ－ＶＯＬ）または副ボリューム（Ｓ－ＶＯＬ）に対し、ＳＣＳＩリザーブコマンドを発行する（１１１）。

【００４１】

ＳＣＳＩリザーブコマンドに対するボリュームからのレスポンスが正常終了で

あった場合（１１２）はボリュームが判定用ペアボリュームの正ボリューム（Ｐ－ＶＯＬ）か副ボリューム（Ｓ－ＶＯＬ）かを調査する為にボリューム制御用ソフトウェア８４は記憶装置に対してボリューム属性確認コマンドを発行する（１１３）。判定用ボリュームが正ボリューム（Ｐ－ＶＯＬ）であった場合は更にペア再同期コマンドを発行する（１１４）。

【００４２】

一方、判定用ボリュームが副ボリューム（Ｓ－ＶＯＬ）であった場合はペアスワップ再同期コマンドを発行する（１１５）。ここで、ペア再同期コマンドとはサスペンド状態のペアボリュームを再同期しペア状態に遷移させるコマンドである。

【００４３】

一方、ペアスワップ再同期コマンドとはサスペンド状態のペアボリュームを再同期しペア状態に遷移させ、かつ正ボリュームを副ボリュームに、副ボリュームを正ボリュームにスワップするコマンドである。これにより、副ボリュームであった判定用ボリュームが正ボリュームに遷移する。

【００４４】

図４に示すように、ペア再同期コマンドはコマンド発行配下のボリュームすなわち自ボリュームが正ボリューム（Ｐ－ＶＯＬ）かつサスペンド状態（正常）の場合のみ成功する（１３０）。

【００４５】

一方、ペアスワップ再同期コマンドは自ボリュームが副ボリューム（Ｓ－ＶＯＬ）かつサスペンド状態（正常）の場合のみ成功する（１３１）。

【００４６】

また、図３に示すように、ペア再同期コマンドまたはペアスワップ再同期コマンドのどちらかが発行され、実行状態になった場合に、ペアボリュームの状態は「サスペンド状態（正常）」（１２０）から「コピー状態」（１２１）さらに、コピーが完了すると「ペア状態」（１２２）に遷移する。よって、図４に示すように、ペアボリュームの状態が「ペア状態」になった後で発行されたペア再同期コマンドまたはペアスワップ再同期コマンドは異常終了することになる。

【００４７】

このようにペアボリューム制御用ソフトウェアはペア再同期またはペアスワップ再同期コマンドが正常終了した場合に、ＳＣＳＩリザーブコマンドが成功したことを上位装置へＧｏｏｄステータス（成功）を返して報告する（１１７）。すなわちクラスタの専用使用権が取れたことを報告する。

【００４８】

一方、ペアボリューム制御コマンドが異常終了した場合は、リザーブに失敗した、すなわちクラスタの専用使用権が取れなかったことを報告する為に、上位装置に対して、Reservation Conflictステータス（失敗）を返す（１１６）。

【００４９】

判定用ボリュームへのリザーブ要求に対して、Ｇｏｏｄステータスを得た上位装置はフェイルオーバを実行し、クラスタ内の全てのリソース、アプリケーション及びサービスを引き継ぐ。

【００５０】

一方、Reservation Conflictステータスを得た上位装置は自身が管理していた全てのリソースを解放し、かつ自身が管理していた全てのアプリケーション及びサービスを停止する。

【００５１】

また図６に示すように、クラスタソフトウェア８０，８１から発行されたＳＣＳＩコマンドがリザーブコマンド以外であった場合はペアボリューム制御用ソフトウェアはペアボリューム制御コマンドを付加せず、そのままの状態で判定用ボリュームに当該ＳＣＳＩコマンドを発行する。

【００５２】

図７から図１３にて本発明をクラスタリングシステムに適用した場合の、システムが正常な状態から障害によりクラスタリングが機能する一連の動作について説明する。まず、最初の状態ではクラスタリングシステムは停止し、判定用ペアボリュームは「サスペンド状態（正常）」を維持している（図７）。

【００５３】

ここで、ユーザがクラスタリングシステムを起動した場合、各上位装置内のク

ラスタソフトウェアはクラスタリングシステム内の専用使用権を獲得する為に、ペアボリューム制御用ソフトウェアに対してＳＣＳＩリザーブコマンドを発行する。ＳＣＳＩリザーブコマンドを受領したペアボリューム制御用ソフトウェアは、判定用ペアボリュームの正ボリューム（判定用（Ｐ））または副ボリューム（判定用（Ｓ））に対し、ＳＣＳＩリザーブコマンドを発行し、専用使用権争いを行う。本例では各システム内に一台の上位装置のみ存在するので、上位装置１０，１１が発行するＳＣＳＩリザーブコマンドに対する各ボリュームからペアボリューム制御用ソフトウェアに対するレスポンスはそれぞれ正常終了となる。

【００５４】

次に、判定用ペアボリュームの正ボリューム（判定用（Ｐ））側に接続するペアボリューム制御用ソフトウェアはペア再同期コマンドを発行する。一方、判定用ペアボリュームの副ボリューム（判定用（Ｓ））側に接続するペアボリューム制御用ソフトウェアはペアスワップ再同期コマンドを発行する。

【００５５】

本例では上位装置１０のペア再同期コマンドが上位装置１１のペアスワップ再同期コマンドより先に発行され、クラスタの専用使用権を獲得したとする。専用使用権を獲得した上位装置１０内のクラスタソフトウェアはWriteコマンドを判定用ペアボリュームに対して発行し、クラスタ管理情報の初期化を行う（図８）。

【００５６】

次に、各上位装置内のクラスタソフトウェアはユーザにて予め設定されたデータボリュームをマウントし、そのデータボリュームに対するアプリケーションを起動する（図９）。

【００５７】

さらに、判定用ペアボリュームに対してはシステムの障害に備え、ペアボリューム制御用ソフトウェアを用いて判定用ペアボリュームを分割することで「サスペンド状態（正常）」に遷移させ（図１０）、システムは通常運用に入る（図１１）。

【００５８】

通常運用中に上位装置の障害又はネットワークの障害等が発生し、ハートビート通信が不可能になった場合、各々の上位装置内のクラスタソフトウェアはクラスタリングシステム内での専用使用権を獲得する為に、ＳＣＳＩバスデバイスリセットを発行し、データ転送インターフェース上をリセットした後、ペアボリューム制御用ソフトウェアに対してＳＣＳＩリザーブコマンドを発行する。ＳＣＳＩリザーブコマンドを受領したペアボリューム制御用ソフトウェアは、判定用ペアボリュームの正ボリューム（判定用（Ｐ））または副ボリューム（判定用（Ｓ））に対し、ＳＣＳＩリザーブコマンドを発行し、専用使用権争いを行う。

【００５９】

本例では各サイト内に一台の上位装置のみ存在するので、上位装置１０，１１が発行するＳＣＳＩリザーブコマンドに対し、各ボリュームからペアボリューム制御用ソフトウェアに対するレスポンスはそれぞれ正常終了となる。次に、判定用ペアボリュームの正ボリューム（判定用（Ｐ））側に接続するペアボリューム制御用ソフトウェアはペア再同期コマンドを発行する。

【００６０】

一方、判定用ペアボリュームの副ボリューム（判定用（Ｓ））側に接続するペアボリューム制御用ソフトウェアはペアスワップ再同期コマンドを発行する。本例では上位装置１０のペア再同期コマンドが上位装置１１より先に発行され、クラスタの専用使用権を獲得したとする（図１２）。

【００６１】

専用使用権を獲得した上位装置１０内のクラスタソフトウェアは他の上位装置のアプリケーション及びディスクボリューム等のリソースを全て引き継ぎ、フェイルオーバを実行する。一方、判定用ボリュームに対するＳＣＳＩリザーブコマンドが失敗した上位装置はその上位装置上で実行していたアプリケーション及びサービスを全て停止する（図１３）。

【００６２】

このように、本発明により分散して設置された２台のリモートコピー機能を持つ記憶装置間で作成されたペアボリュームを、１台の上位装置が当該ペアボリュームを排他的に占有し、他の上位装置からの更新要求を拒否することで、ペアボ

リュームを当該複数の上位装置に対して１つのボリュームとして認識させることができ、当該ペアボリュームをクラスタリングシステムの判定用ボリュームに適用することができる。

【００６３】

【発明の効果】

本発明は分散して設置された２台のリモートコピー機能を持つ記憶装置間で作成されたペアボリュームを、２台以上の上位装置から共用する構成において、１台の前記上位装置が前記ペアボリュームを排他的に占有し、他の前記上位装置からの更新要求を拒否することで、前記ペアボリュームを前記複数の上位装置に対して１つのボリュームとして認識させるデータ格納システムを実現する。

【図面の簡単な説明】

【図１】

本発明の一実施形態を示すハードウェア構成図である。

【図２】

リモートコピー機能のペアボリュームを判定用ボリュームに使用したクラスタリングシステムの構成図である。

【図３】

ペアボリュームの状態遷移図である。

【図４】

ペア再同期コマンド又はペアスワップ再同期コマンド発行する場合の、各ボリューム属性に対する記憶装置からのレスポンスを示す

【図５】

ＳＣＳＩリザーブコマンド受領時のペアボリューム制御用ソフトウェアの動作である。

【図６】

ＳＣＳＩリザーブコマンド以外受領時のペアボリューム制御用ソフトウェアの動作である。

【図７】

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

【図８】

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

【図９】

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

【図１０】

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

【図１１】

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

【図１２】

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

【図１３】

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

【符号の説明】

１０、１１上位装置

１２，１３クラスタソフトウェア

２０、２１記憶装置

２２　正ボリューム（Ｐ－ＶＯＬ）

２３　副ボリューム（Ｓ－ＶＯＬ）

３０，３１　データ転送用インターフェース

４０　リモートコピー用インターフェース

５０　ハートビート用ネットワーク

６０　判定用ペアボリュームの正ボリューム（Ｐ－ＶＯＬ）

６１　判定用ボリュームの副ボリューム（Ｓ－ＶＯＬ）

７０、７３　データボリュームの正ボリューム（Ｐ－ＶＯＬ）

７１、７２データボリュームの副ボリューム（Ｓ－ＶＯＬ）

８０，８１　クラスタソフトウェア

８２，８３　アプリケーション

８４，８５　ペアボリューム制御用ソフトウェア

９０，９１　標準デバイスドライバ

１００、１０１　サイト

１１０　クラスタソフトウェアのＳＣＳＩリザーブコマンド発行

１１１　ペアボリューム制御用ソフトウェアのＳＣＳＩリザーブコマンド発行

１１２　実行結果の判定

１１３　ボリューム属性の確認

１１４　ペア再同期コマンド

１１５　ペアスワップ再同期コマンド

１１６　リザーブコマンドの失敗（Reservation Conflictステータス）１１７　リザーブコマンドの成功（Ｇｏｏｄステータス）

１２０　サスペンド状態（正常）

１２１　コピー状態

１２２　ペア状態

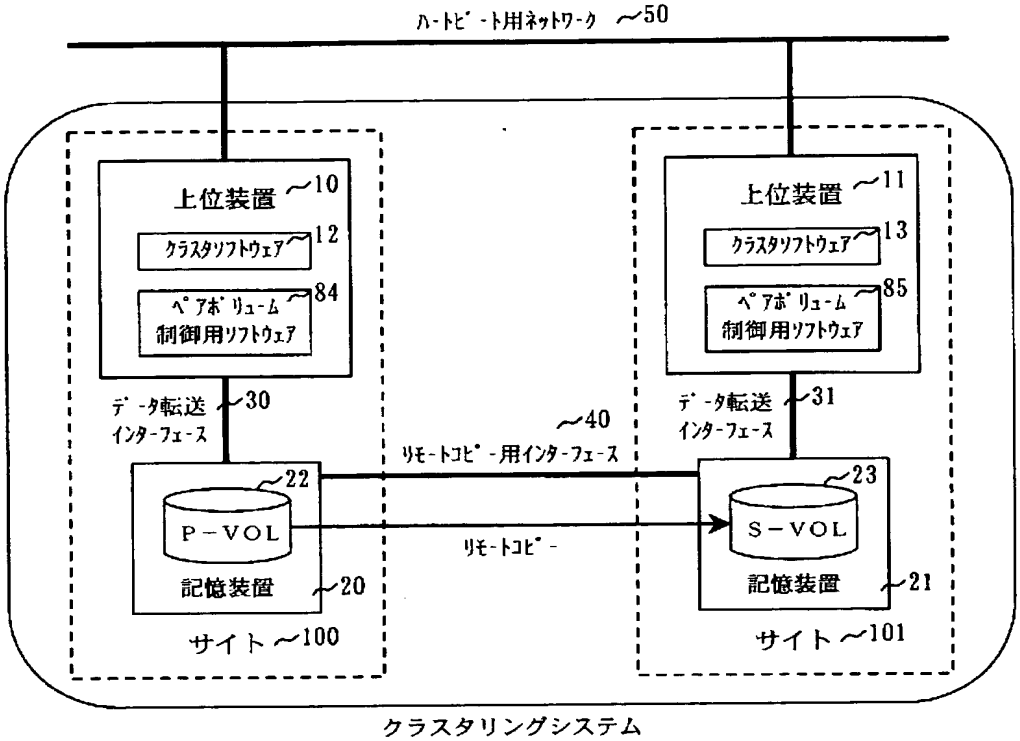
１３０　ペア再同期コマンドの正常終了

１３１　ペアスワップ再同期コマンドの正常終了

【書類名】　図面

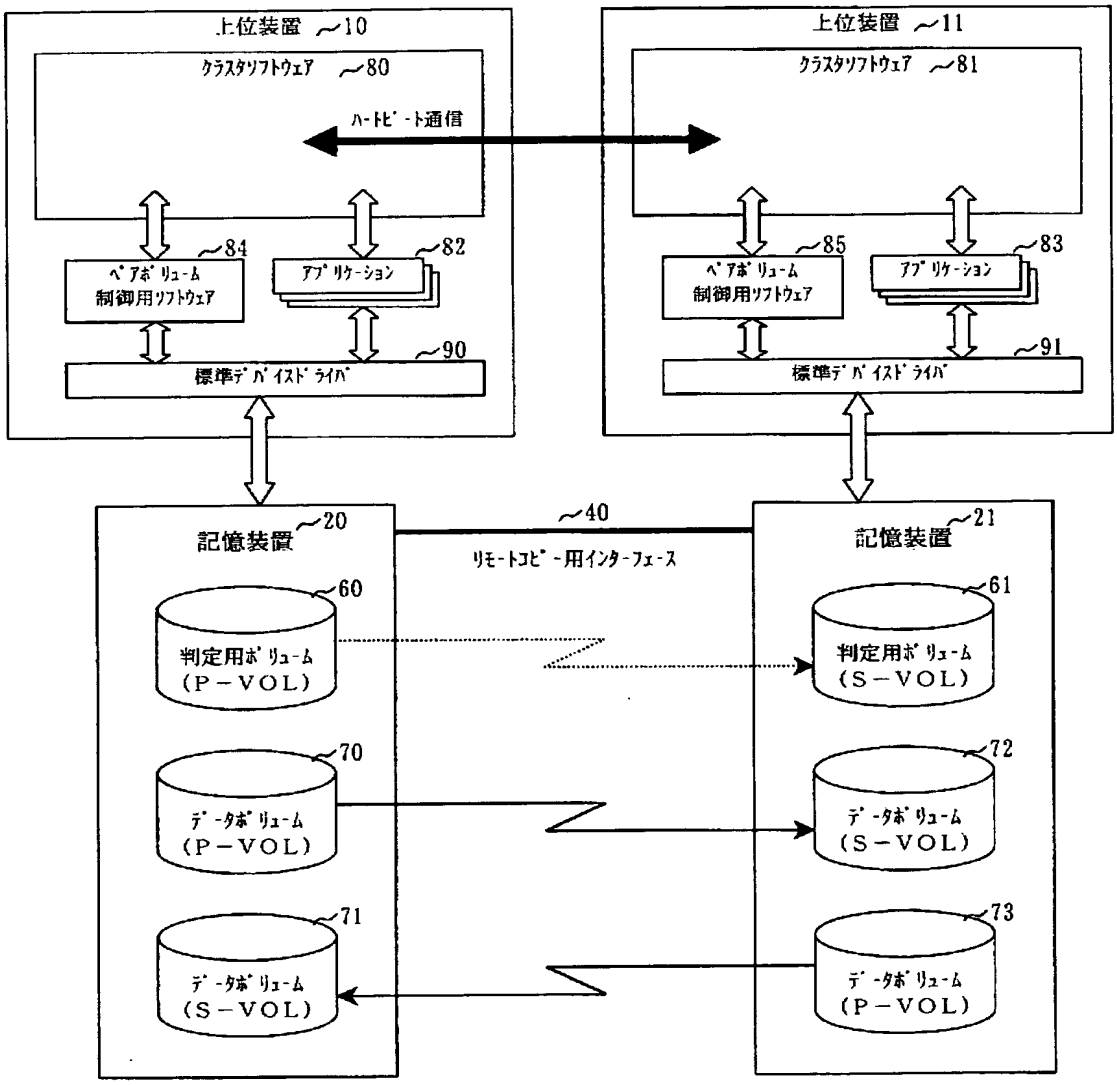
【図１】

図 1


ハートビート用ネットワーク
50
上位装置
10
クラスタソフトウェア
12
ペアボリューム制御用ソフトウェア
84
データ転送インターフェース
30
22
P－VOL
記憶装置
20
サイト
100
40
リモートコピー用インターフェース
リモートコピー
上位装置
11
クラスタソフトウェア
13
ペアボリューム制御用ソフトウェア
85
データ転送インターフェース
31
23
S－VOL
記憶装置
21
サイト
101
クラスタリングシステム

【図２】

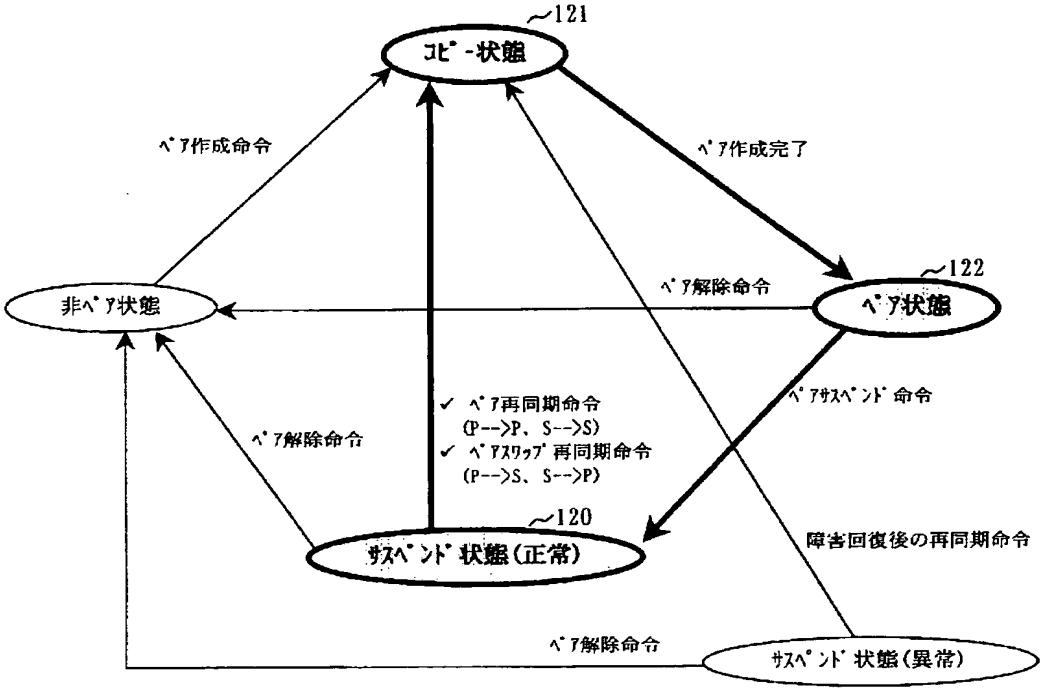
図 2


上位装置 ～10
クラスタソフトウェア ～80
ハートビート通信
ペアボリューム制御用ソフトウェア
～84
アプリケーション
～82
標準デバイスドライバ
～90
上位装置 ～11
クラスタソフトウェア ～81
ペアボリューム制御用ソフトウェア
～85
アプリケーション
～83
標準デバイスドライバ
～91
記憶装置 ～20
～40
リモートコピー用インターフェース
記憶装置 ～21
判定用ボリューム (P－VOL)
～60
判定用ボリューム (S－VOL)
～61
データボリューム (P－VOL)
～70
データボリューム (S－VOL)
～72
データボリューム (S－VOL)
～71
データボリューム (P－VOL)
～73

【図３】

図 3

ペアボリュームの状態遷移図



【図４】

図 4

ペア再同期コマンド発行時の記憶装置からのレスポンス

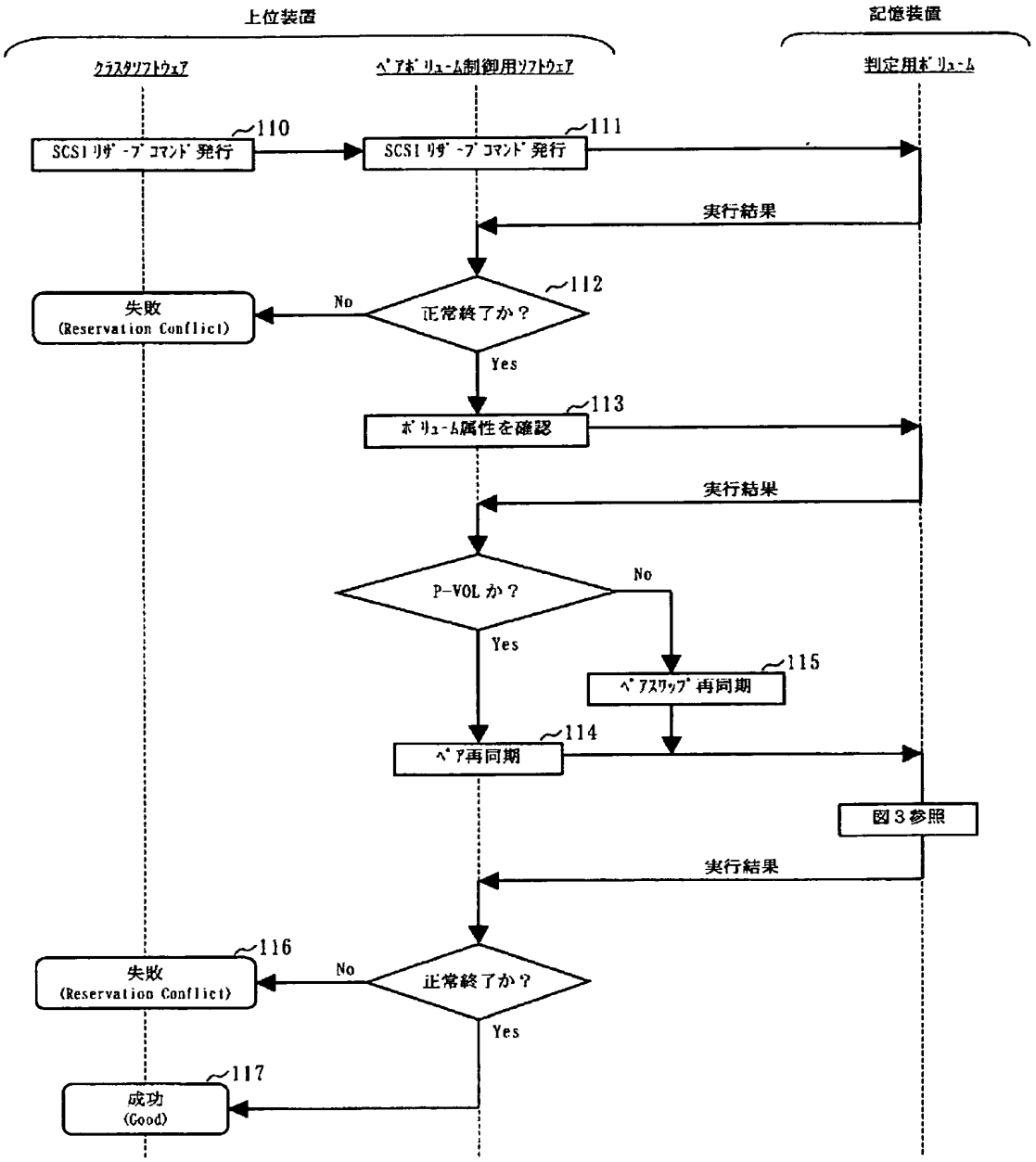
実行ｺﾏﾝﾄﾞ	自ﾎﾞﾘｭｰﾑの属性		記憶装置のﾚｽﾎﾟﾝｽ
ﾍﾟｱ再同期	非ﾍﾟｱﾎﾞﾘｭｰﾑ		異常終了
	P-VOL	ｺﾋﾟｰ中	異常終了
		ﾍﾟｱ状態	異常終了 ～130
		ｻｽﾍﾟﾝﾄﾞ状態(正常)	正常終了
		ｻｽﾍﾟﾝﾄﾞ状態(異常)	異常終了
	S-VOL	ｺﾋﾟｰ中	異常終了
		ﾍﾟｱ状態	異常終了
		ｻｽﾍﾟﾝﾄﾞ状態(正常)	異常終了
		ｻｽﾍﾟﾝﾄﾞ状態(異常)	異常終了
ﾍﾟｱｽﾜｯﾌﾟ再同期	非ﾍﾟｱﾎﾞﾘｭｰﾑ		異常終了
	P-VOL	ｺﾋﾟｰ中	異常終了
		ﾍﾟｱ状態	異常終了
		ｻｽﾍﾟﾝﾄﾞ状態(正常)	異常終了
		ｻｽﾍﾟﾝﾄﾞ状態(異常)	異常終了
	S-VOL	ｺﾋﾟｰ中	異常終了
		ﾍﾟｱ状態	異常終了 ～131
		ｻｽﾍﾟﾝﾄﾞ状態(正常)	正常終了
		ｻｽﾍﾟﾝﾄﾞ状態(異常)	異常終了

【図５】

図５

ペアボリューム制御用ソフトウェアの動作

［リザーブコマンド］

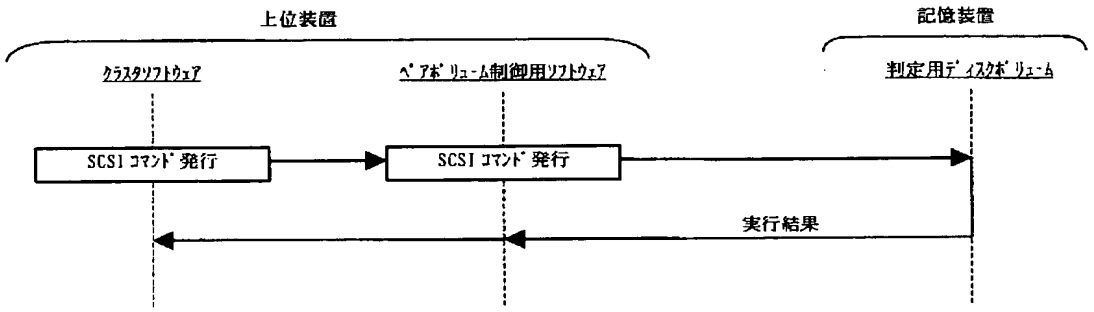


【図６】

図 6

ペアボリューム制御用ソフトウェアの動作

[リザーブコマンド以外]

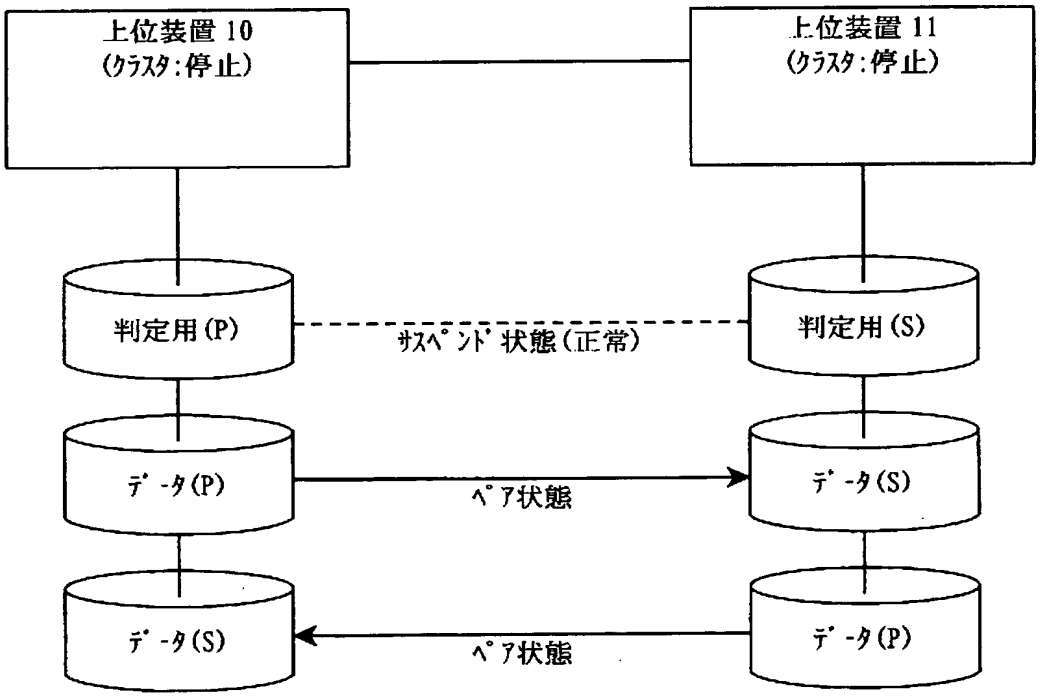


【図７】

図 7

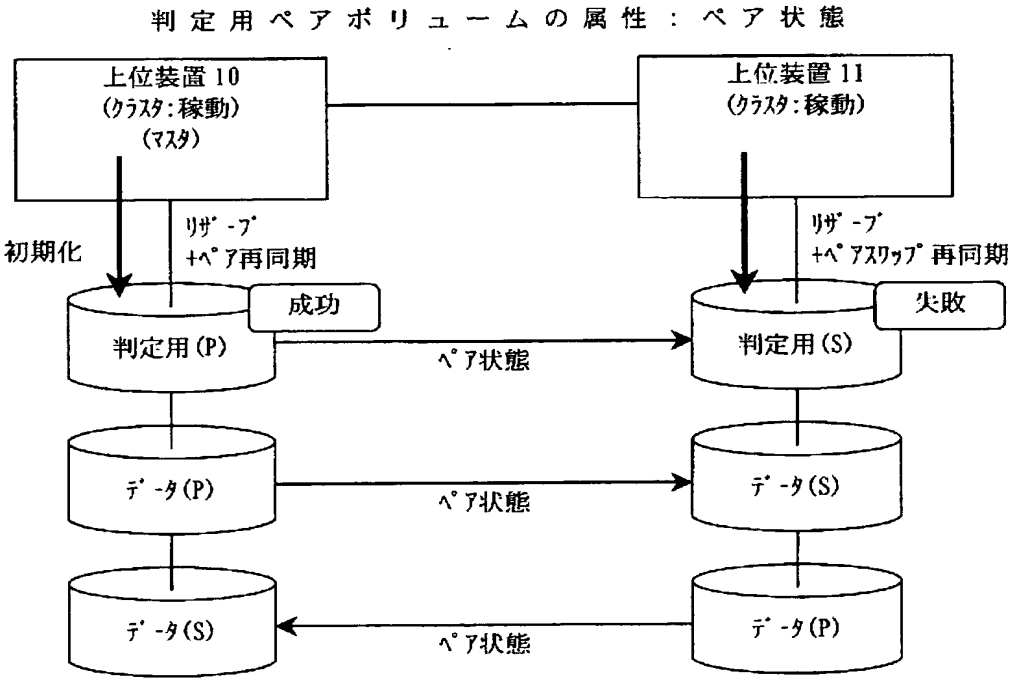
初期状態

判定用ペアボリュームの属性：サスペンド（正常）



【図８】

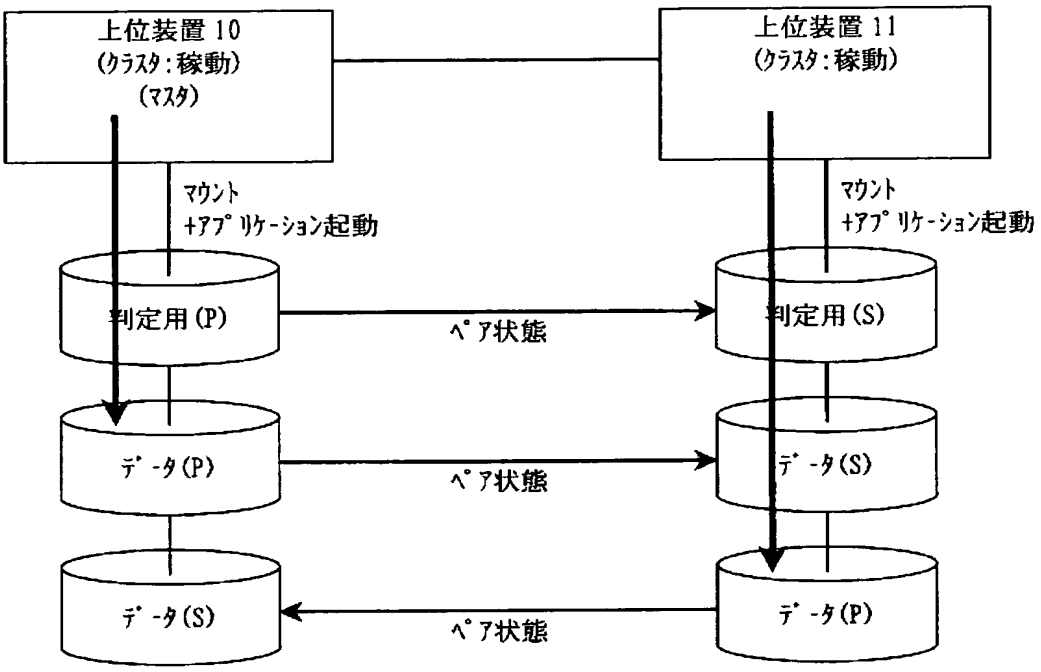
図 8


判定用ペアボリュームの属性：ペア状態
上位装置 10
(ｸﾗｽﾀ:稼動)
(ﾏｽﾀ)
上位装置 11
(ｸﾗｽﾀ:稼動)
初期化
ﾘｻﾞｰﾌﾞ
+ﾍﾟｱ再同期
ﾘｻﾞｰﾌﾞ
+ﾍﾟｱｽﾜｯﾌﾟ再同期
成功
失敗
判定用(P)
ﾍﾟｱ状態
判定用(S)
ﾃﾞｰﾀ(P)
ﾍﾟｱ状態
ﾃﾞｰﾀ(S)
ﾃﾞｰﾀ(S)
ﾍﾟｱ状態
ﾃﾞｰﾀ(P)

【図９】

図 9

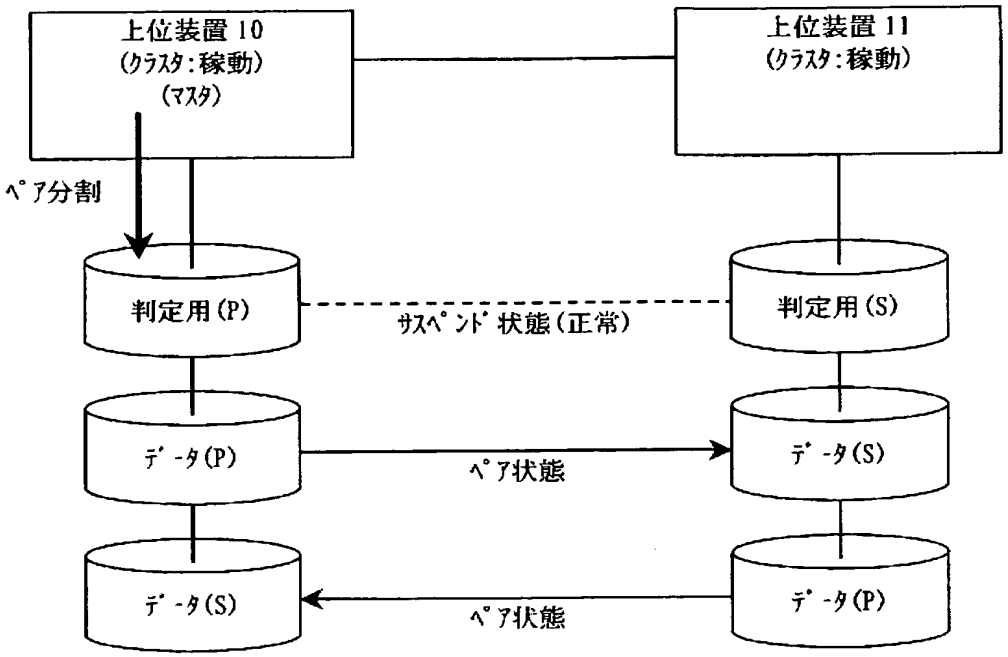
判定用ペアボリュームの属性：ペア状態



【図１０】

図１０

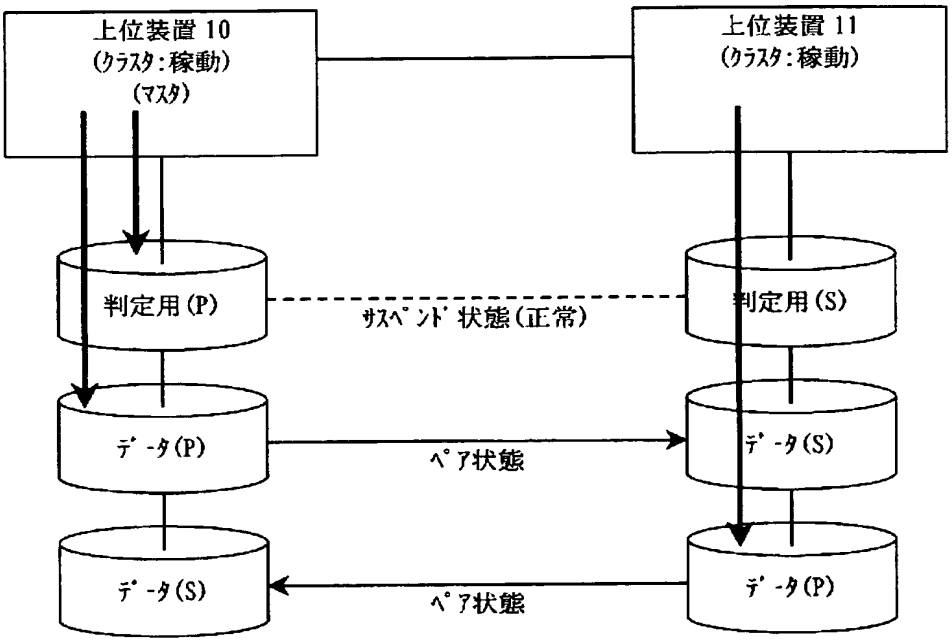
判定用ペアボリュームの属性：サスペンド（正常）



【図１１】

図１１

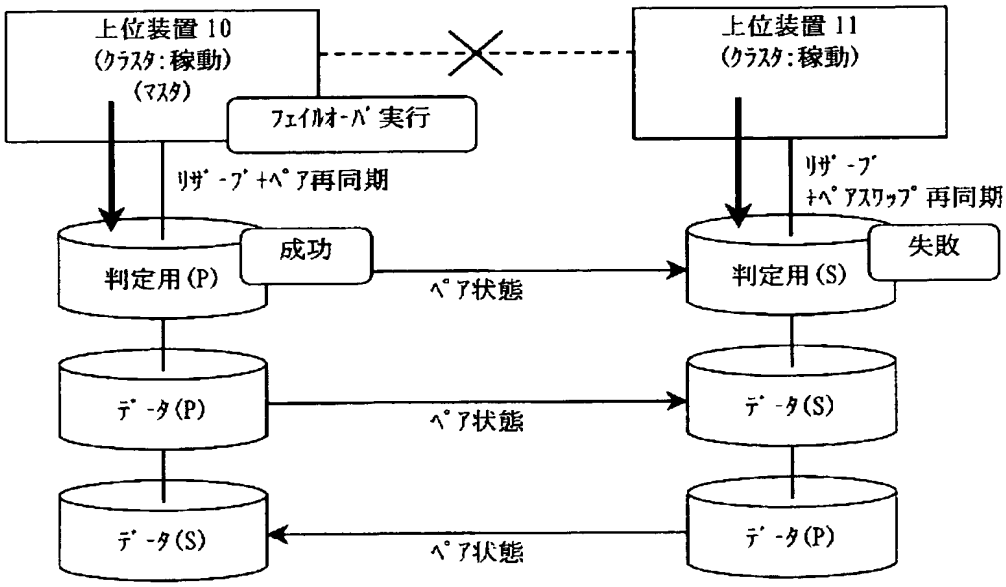
判定用ペアボリュームの属性：サスペンド（正常）



【図１２】

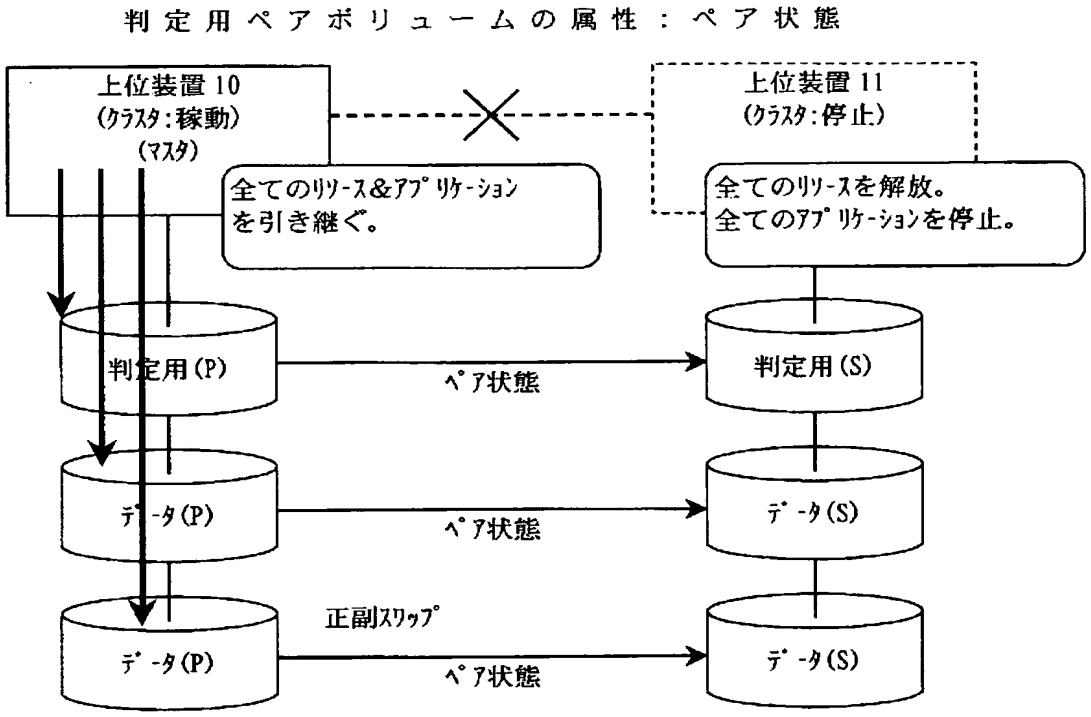
図１２

判定用ペアボリュームの属性：ペア状態



【図１３】

図１３


判定用ペアボリュームの属性：ペア状態
上位装置10
(ｸﾗｽﾀ:稼動)
(ﾏｽﾀ)
全てのﾘｿｰｽ&ｱﾌﾟﾘｹｰｼｮﾝ
を引き継ぐ。
上位装置11
(ｸﾗｽﾀ:停止)
全てのﾘｿｰｽを解放。
全てのｱﾌﾟﾘｹｰｼｮﾝを停止。
判定用(P)
ﾍﾟｱ状態
判定用(S)
ﾃﾞｰﾀ(P)
ﾍﾟｱ状態
ﾃﾞｰﾀ(S)
正副ｽﾜｯﾌﾟ
ﾃﾞｰﾀ(P)
ﾍﾟｱ状態
ﾃﾞｰﾀ(S)

【書類名】　要約書

【要約】

【課題】

リモートコピー機能によって生成されたペアボリュームを、クラスタリングシステムの判定用ボリュームとして使用するデータ格納システムを実現する。

【解決手段】

分散して設置された２台のリモートコピー機能を持つ記憶装置２０，２１間で作成されたペアボリューム２２，２３を、２台以上の上位装置１０，１１から共用する構成において、上位装置１０、１１上のクラスタソフトウェア１２，１３から発行されるリザーブコマンド等のＳＣＳＩコマンドを一旦、ペアボリューム制御用ソフトウェア８４，８５が受け、当該ペアボリュームが唯一で共通の「状態」となるようにして操作するペアボリューム制御コマンドとを、リザーブコマンド等のＳＣＳＩコマンドと組み合わせて発行することにより、ペアボリュームを排他的使用する上位装置を一意に決定する。

【効果】

リモートコピー機能によって生成されたペアボリュームを、クラスタリングシステムの判定用ボリュームとして使用することが可能となる。

【選択図】　図１

認定・付加情報

特許出願の番号	特願２００１－２６６６２９
受付番号	５０１０１２９２８４８
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１３年　９月　５日

＜認定情報・付加情報＞

【提出日】　平成13年 9月 4日

次頁無

特願２００１－２６６６２９

出　願　人　履　歴　情　報

識別番号　［０００００５１０８］

１．変更年月日　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所

２．変更年月日　２００４年　９月　８日
［変更理由］　住所変更
住　所　東京都千代田区丸の内一丁目６番６号
氏　名　株式会社日立製作所